METHODOLOGICAL REVIEW ARTICLE

Modeling Systematicity and Individuality in Nonlinear Second Language Development: The Case of English Grammatical Morphemes

Akira Murakami

University of Cambridge

This article introduces two sophisticated statistical modeling techniques that allow researchers to analyze systematicity, individual variation, and nonlinearity in second language (L2) development. Generalized linear mixed-effects models can be used to quantify individual variation and examine systematic effects simultaneously, and generalized additive mixed models allow for the examination of systematicity, individuality, and nonlinearity within a single model. Based on a longitudinal learner corpus, this article illustrates the usefulness of these models in the context of L2 accuracy development of English grammatical morphemes. I discuss the strengths of each technique and the ways in which these techniques can benefit L2 acquisition research, further highlighting the importance of accounting for individual variation in modeling L2 development.

Keywords statistical modeling; mixed-effects model; generalized additive mixed model; learner corpus; individual variation; grammatical morphemes

Author Note

I would like to express my sincere gratitude to Dora Alexopoulou, who as my supervisor provided me with continuing and valuable guidance, extensive feedback, and constructive advice on my PhD project that this article is based on. I would also like to thank John Williams and Detmar Meurers for their comments on my PhD dissertation.

This article has been awarded an Open Data badge. All data are publicly accessible via the Open Science Framework at https://osf.io/dbuh4. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Akira Murakami, Department of Theoretical and Applied Linguistics, 9 West Road, University of Cambridge, Cambridge, United Kingdom, CB3 9DP. E-mail: am933@cam.ac.uk

Introduction

For many years, second language acquisition (SLA) research has focused on revealing systematicity in second language (L2) development. The problem with searching for systematicity alone is that the identification of systematic patterns often necessitates statistical averaging, and averaging conceals individual patterns (Dörnyei, 2009). Indeed, in psychology, it is well-known that the averaged pattern can differ from the individual patterns that constitute the data (e.g., Heathcote, Brown, & Mewhort, 2000). Therefore, there has recently been a growing interest in SLA in understanding the performance of individual learners (van Geert & van Dijk, 2002; Verspoor, Lowie, & van Dijk, 2008).

However, studying individual variation requires appropriate analytical tools.

Conventional statistical techniques in SLA, such as analyses of variance (ANOVAs), cannot appropriately disentangle between- and within-learner variability. With recent developments in statistical modeling, however, we can now model and analyze group- and individual-level features simultaneously. The technique, called mixed-effects modeling, is now widely used in (applied) linguistics, including SLA (e.g., Kozaki & Ross, 2011; Tremblay, Derwing, Libben, & Westbury, 2011; see also Cunnings, 2012, and Linck & Cunnings, 2015). In most studies employing mixed-effects models, however, the technique has been used to control for individual differences in testing the significance of predictors or to study the sources of these differences. While this is certainly useful, mixed-effects models can also provide information about the amount and pattern of individual variation (Dingemanse & Dochtermann, 2013; Kliegl, Wei, Dambacher, Yan, & Zhou, 2011). This article not only reports on tests of the effect of predictors but also is focused on individual variation disclosed by mixed-effects models, with the aim of demonstrating that mixed-effects analysis can model systematicity and individuality simultaneously.

Nonlinearity in SLA

Another recent trend in SLA research is the emphasis on the process rather than the product of learning (Atkinson, 2011). The learning process, however, is never linear. There is ample empirical evidence to demonstrate nonlinearity in L2 development. Perhaps the best known examples of nonlinearity include U-shaped development (e.g., Lightbown, 1983) and power-law development (e.g., DeKeyser, 1997; Ellis & Schmidt, 1998). In U-shaped development, accuracy is high at the beginning, and it temporarily decreases before becoming high again. In power-law development, decrement in error becomes progressively smaller as the learner develops. Because power-law development covers the entire span of development and does not exhibit a systematic

decrease in accuracy in the process, U-shaped and power-law development are mutually exclusive.

Despite the prevalence of nonlinearity in L2 development, researchers are not fully equipped with appropriate statistical tools to analyze it. Classical statistical analysis is generally incapable of analyzing the learning process, including nonlinearity (Larsen-Freeman, 2011; see also Baayen, 2010b). For instance, if researchers want to investigate the effect of a treatment on the linguistic complexity of learners' writing while controlling for their proficiency, there is not sufficient evidence to assume a particular functional form between proficiency and linguistic complexity; thus, it is not straightforward to statistically control proficiency. As in individual differences analysis, however, recent development in statistics allows researchers to model nonlinearity. Although this technique—the generalized additive model (Hastie & Tibshirani, 1990)—is new to SLA, it has been used in other areas of (applied) linguistics including psycholinguistics (e.g., Baayen, 2010a; Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011) and sociolinguistics (Wieling, Nerbonne, & Baayen, 2011). This article illustrates its usefullness for SLA.

Aim and Research Questions

The aim of this article is to introduce to the SLA community two statistical modeling techniques that take into account systematicity, individual variation, and nonlinearity. I do so by modeling the development of accuracy of English grammatical morphemes in L2 learners. Grammatical morphemes were targeted in this exposition because their acquisition has been extensively studied in SLA since its early days (e.g., Dulay & Burt, 1973), and much is already known about the variables that affect their accuracy. This allows the focus to be on what the new techniques can contribute to the field.

The status of the morpheme has been challenged as a functional unit of representation (e.g., Baayen et al., 2011; Ellis & Schmidt, 1998; Plaut & Gonnerman, 2000). Bybee (1985, 2010), for instance, demonstrated the essentially gradient nature of grammatical morphemes. In the process of grammaticalization (e.g., the lexical item *go* came to be used as an auxiliary verb in the future construction of *be going to*; Bybee, 2010), for example, the meaning and function of a lexical item undergo a gradual change without a clear-cut boundary between a lexical and a grammatical item. The historical account further shows that the word, and not the morpheme, has been regarded as the smallest unit of a grammatical system (Blevins, 2013). Given the methodological focus of this article, however, this issue is rather marginal. In this article, two research questions are posed:

1. How large is individual variation in the developmental pattern of morphemes?
2. Do their cross-sectional and longitudinal developmental patterns vary depending on the particular morpheme and on whether learners' native languages (L1s) have an equivalent morpheme?

The background of the first question is that, while SLA has identified prototypical developmental patterns, individual learners are hypothesized to exhibit a variety of learning curves. Therefore, I investigate the extent to which individual variation is observed in the developmental patterns of morphemes. With regard to the second question, in addition to individual variation, I address the systematic effect of L1, which is known to affect nearly every aspect of L2 development (Jarvis & Pavlenko, 2007; Odlin, 1989), including grammatical morphemes (Luk & Shirai, 2009; Murakami & Alexopoulou, 2015). It is not clear, however, how L1 influence emerges or changes during the acquisitional process (Jarvis, 2000).

I further investigate whether the developmental pattern differs across morphemes. Prior

research has often drawn distinctions between free versus bound and verbal versus nominal morphemes (e.g., Brown, 1973; Goldschneider & DeKeyser, 2001). Slobin (1996) further distinguished between the morphemes that encode language-independent concepts (e.g., number as expressed by plural *–s*) and those that encode language-dependent concepts (e.g., definiteness as expressed by articles). It would, therefore, be natural to observe differences in developmental patterns between morphemes as well. By modeling both systematicity and individuality simultaneously, I aim to demonstrate a more comprehensive view of morpheme accuracy development.

The analyses featured here do not presuppose knowledge of generalized additive (mixed) models. It is assumed, however, that readers are familiar with the basic ideas of regression modeling, including generalized linear models and model comparison based on information-theoretic measures, such as Akaike information criterion (AIC). It is further assumed that readers have basic knowledge of mixed-effects models. Appendix S1 in the Supporting Information online provides an introduction to general ideas in regression modeling necessary for this article.

Data Source and Analysis

Corpus

For this study, I employed the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou, & Korhonen, 2014), which is publicly available at http://corpus.mml.cam.ac.uk/efcamdat. The learner corpus includes writings from Englishtown, an online school run by Education First. A course in Englishtown consists of 16 levels, with eight units each. Learners usually progress from lower to higher levels unit by unit, although they are free to go back or skip units. A placement test suggests an appropriate level at which learners begin their coursework. At the end of each unit is a free writing task on a variety of

topics (e.g., shopping, writing an email). A model answer is provided for each writing task, and learners can consult the sample and other external resources, such as dictionaries in the process of writing. Each writing task specifies length, with assignments ranging from 20 to 40 words in Level 1 Unit 1 to 150 to180 words in Level 16 Unit 8. Teachers provide feedback on writing, including the correction of erroneous grammatical morphemes. The present study used teacher feedback as error tags, which were exploited to calculate accuracy. Error tags are not annotated in all of the writings, however. Apart from learners' writings, EFCAMDAT includes, for each text, such metadata as the ID of the learner, his or her country of residence, and the date and time of submission. This information allows researchers to track the longitudinal development of individual learners.

Target Morphemes

The initial set of targets included six English grammatical morphemes: articles, past tense *–ed*, plural *–s*, possessive *–s*, progressive *–ing*, and third person singular *–s*. These are the morphemes that have often been targeted in SLA literature (cf. Goldschneider & DeKeyser, 2001). However, possessive *–s* was dropped because it did not occur frequently enough to allow for the investigation of individual variation or longitudinal development. Furthermore, progressive *–ing* and third person *–s* were dropped because their accuracy was close to 100% throughout learners' development. High accuracy rates made the inspection of development difficult because learners who barely achieved 100% accuracy could not be distinguished from those who did so effortlessly (i.e., the ceiling effect). Thus, the final set of target morphemes was composed of articles, past tense *–ed*, and plural *–s*. Articles included both definite and indefinite articles. Past tense *–ed* included only regular past tense forms (e.g., *opened*) and not irregular ones (e.g., *thought*). Similarly, plural *–s* included only regular forms (e.g., *cups*) and not irregular ones (e.g.,

mice).

Target L1 Groups and Proficiency Levels

The current analyses focus on the following 10 L1 groups with the largest amount of data in EFCAMDAT: Brazilian Portuguese, Mandarin Chinese, German, French, Italian, Japanese, Korean, Russian, Spanish, and Turkish. As EFCAMDAT does not provide direct information about learners' L1s, such information was inferred from the countries in which learners resided, providing a close approximation[1]. L1 Mandarin Chinese learners included those living in Mainland China and in Taiwan, and L1 Spanish learners included those living in Spain and Mexico. L1 Mandarin Chinese is referred to as L1 Chinese and L1 Brazilian Portuguese as L1 Brazilian to save space. The Englishtown proficiency levels 1–3, 4–6, 7–9, 10–12, 13–15, and 16 correspond to levels A1 through C2 in the Common European Framework of Reference (CEFR).

Subcorpus

The study only included learners whose sum of obligatory contexts and overgeneralization errors in error-tagged texts was 10 or more for each of the three morphemes. In addition, due to the high computational cost of some analyses, it was necessary to limit the data to a maximum of 20 learners from each L1 group. The 20 learners selected were those with the largest number of writings within the L1 group. Because the L1 French, Japanese, Korean, and Turkish groups included 20 or fewer learners after applying the first selection criterion (i.e., obligatory contexts plus overgeneralization errors $\geq$ 10), the second criterion was not relevant to these four groups. Figure 1 shows the distribution of learners and error-tagged writings across L1 groups and Englishtown levels. Learner level was operationalized as the learner's mean level in Englishtown. In all, there were 3,323 writings from 158 learners. The subcorpus included 315,141 words in total, and the mean number of words per writing was 94.8 (*SD* = 50.0).

FIGURE 1

Accuracy Measure and Data Extraction

This study employed the ratio between correct uses and errors as a measure of accuracy. The number of correct uses was obtained by subtracting the number of omission and misformation errors from that of obligatory contexts. Obligatory contexts were operationalized as morpheme use in the corrected text, that is, the text in which incorrect portions were replaced with the corresponding corrected forms based on error tags. For instance, if a learner wrote, *She has a big nose and small mouth*, and it was corrected to read, *She has a big nose and a small mouth*, there were two obligatory contexts for articles because the article occured twice in the corrected sentence. The number of errors was the sum of omission, misformation, and overgeneralization errors. This accuracy measure is conceptually equivalent to targetlike use (TLU) scores (Pica, 1983). In visualizing accuracy, the study used TLU scores, which are calculated by dividing the number of correct uses by the sum of the numbers of obligatory contexts and overgeneralization errors. R scripts were written to count the frequency of obligatory contexts and each type of error in error-tagged texts. The accuracy of the R scripts is reported in Appendix S2 in the Supporting Information online.

The information provided by error annotation leads to intriguing insights into patterns of accuracy development, although the use of teacher feedback as error annotation can also introduce noise to the data. A manual given to Englishtown teachers asks them to be complete in providing feedback, and it explicitly mentions articles, plural *–s*, and verb tense among the features teachers should pay attention to. This briefing should raise the accuracy and comprehensiveness of error annotation.

Variables and Analysis

Accuracy was modeled as a function of several variables, and models were compared to address the research questions. The dependent variable was accuracy in the form of odds. In the variants of logistic regression models employed in this study, the number of correct uses was entered as the number of successes, and the number of errors was entered as the number of failures. There were four independent variables: proficiency, writing number (writingnum), morpheme, and L1 type (L1type).

- Proficiency was represented by the mean Englishtown level calculated from the level and unit at which the learner submitted his or her writings. The value was unchanged within learners, and the variable was meant to capture between-learner, cross-sectional development. Proficiency was standardized to facilitate interpretation. The mean and standard deviation of proficiency were 51.8 (Level 7 Unit 4) and 22.6, respectively.
- Writing number represented the within-learner writing order. One indicated the first writing of a learner, two indicated the second writing, and so forth. Writing numbers were assigned to both error-tagged and untagged writings so that development over untagged writings could be interpolated. This variable was meant to capture within-learner, longitudinal development, and was standardized over learners after its values were centered within each learner. Accordingly, zero in the standardized writing number indicated the mean writing number within each learner. The standard deviation of the writing number was 15.6.
- Morpheme was a categorical variable with three levels: one for each morpheme, with articles as the reference level.
- L1 type was a dichotomous variable representing L1 influence and indicating whether a

L1 had an equivalent morpheme. The L1 type had two levels: ABSENT and PRESENT. The ABSENT group was the reference level. The ABSENT group included L1 groups who lacked the equivalent linguistic features in their L1s, and L1 groups for whom the marking of the feature is optional. By contrast, the PRESENT group must mark equivalent features. For instance, L1 Japanese was considered to be in the ABSENT group for the article because it is not obligatory in Japanese to express definiteness, the central concept of the English article system. Conversely, Japanese was considered to belong to the PRESENT group in past tense *–ed* because the Japanese morpheme *–ta* roughly corresponds to past tense *–ed* in English, and it is difficult to express pastness without the use of this morpheme in Japanese. This approach to representing the effect of L1 is rather crude and oversimplified, but as will be shown, it is useful for capturing L1 influence (Murakami & Alexopoulou, 2015). The ABSENT group included L1 Chinese, Japanese, Korean, Russian, and Turkish for articles; L1 Chinese for past tense *–ed*; and L1 Chinese, Japanese, and Korean for plural *–s*. The remaining L1s were included in the PRESENT group.

In addition to these variables, some of their interactions were entered into the model as well in a stepwise manner. Treatment contrasts were used for categorical variables throughout this article.

Before running the analyses, observations without any obligatory contexts or overgeneralization errors were removed. There were 7,247 nonzero observations across the three morphemes. Table 1 shows the mean number and standard deviation of nonzero observations, obligatory contexts, omission errors, and overgeneralization errors for learners. Naturally, the data size was larger for articles and for plural *–s* than for past tense *–ed* due to the higher frequency of these two morphemes. All of the statistical analyses were performed with R

(version 3.2.1; R Core Team, 2015; cf. Mizumoto & Plonsky, 2015). The R codes and data used here are available via the Open Science Framework at https://osf.io/dbuh4. Before moving on to the main analysis, a cross-sectional view of the data is presented.

TABLE 1

Cross-Sectional View of Morpheme Development

Figure 2 illustrates the cross-sectional development of the three morphemes across L1 types. Each line shows the cross-sectional development in each L1 type. Unlike typical cross-sectional data, however, a learner contributed multiple data points to the figure as he or she produced multiple writings. C2 level was dropped out of the figure due to its small data size but was included in statistical modeling. The fluctuation of accuracy in the graph, which is partially due to the small data size of several observations, makes the close examination of the data difficult. In the following analyses, variants of logistic regression models were employed; each observation was weighed according to its data size in order to investigate whether a significant difference in the developmental pattern across groups could be observed and to determine the extent to which individual variation was present in the development.

FIGURE 2

Taking into Account Individual Variation: Generalized Linear Mixed-Effects Models

The goal of this analysis was to determine the extent to which developmental accuracy patterns varied across individual learners. To quantify individual variation, I employed a generalized linear mixed-effects model (GLMM). Mixed-effects models can handle both systematicity and individuality because they can deal not only with usual within- and between-learner fixed-effects variables, such as morphemes and proficiency (i.e., systematicity), but also with remaining variance across and within learners (i.e., individuality). Partly for this reason, mixed-effects

models have been widely used in longitudinal data analysis (Long, 2012), including in SLA (Barkaoui, 2014; Kozaki & Ross, 2011).

Model Specification and Model Selection

I employed a mixed-effects logistic regression model to analyze the relationship between accuracy, proficiency, longitudinal development, and morpheme. The model included L1 and learner as random-effects factors. Writings were nested within individual learners, who were in turn nested within L1 groups. The model thus had a nested random-effects structure where variance was partitioned into between-L1, between-learner, and between-writing levels (cf. Gries, 2015). Although it would have been possible to construct yet another level by viewing data points as nested within writings, this was not attempted in order to avoid further complexity of the model. By-L1 random intercepts allowed overall accuracy to vary across L1 groups. Variables can also be entered as random contrasts and random slopes. By-L1 random contrasts and random slopes, however, were not entered because the small number of L1 levels (10) might have resulted in unstable models.

The role of each random-effects parameter was as follows. When the by-morpheme random contrasts were present, the by-learner random intercepts allowed article accuracy to vary across individual learners. By-morpheme random contrasts represented individual variation in the accuracy difference between morphemes. The by-writingnum random slope similarly represented individual variation in morpheme-independent learning rates, implying that some writers naturally learned more quickly than others. I was interested in the extent to which I could observe such individual differences and whether—and to what extent—systematic variables (e.g., proficiency) could account for these differences.

I constructed multiple models and found the most plausible model by comparing them.

There has been no agreement on how best to perform model selection in mixed-effects modeling (Gries, 2013). It has been suggested that researchers should use the maximal model, or the model with all possible predictors and the largest random-effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013; Gelman & Hill, 2007; see Baayen, Vasishth, Bates, & Kliegl, 2015, and Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2015, for counter-arguments). However, because the present model had only 10 L1 groups, I opted for an approach where initially the simplest model was built and predictors were added to the model one at a time only if it improved the model. More specifically, the following forward selection approach was used (cf. James, Witten, Hastie, & Tibshirani, 2013). I first built the so-called unconditional model (Bates, 2010) that only included by-L1 and by-learner random intercepts but no fixed-effects predictor. I then added, one by one, a predictor that decreased the model's AIC the most, repeating the procedure until no predictor could improve the model further (but see Whittingham, Stephens, Bradbury, & Freckleton, 2006, for criticisms of the stepwise approach in general). Interaction terms were considered only when the model already included the main effects constituting the interactions. Random contrasts were considered only when the variable was already in the fixed-effects component of the model. Although restricted maximum likelihood (REML) procedures are often used for linear mixed-effects models, I employed maximum likelihood estimation because REML does not allow for the comparison of models with different fixed-effects structure (Bolker, Brooks, Clark, Geange, Poulsen, Stevens, & White, 2009) and also because REML estimates are not well-defined for GLMMs (Bates, 2009). All of the statistical analyses in this section were carried out with the lme4 package (version 1.1-8; Bates, Mächler, Bolker, & Walker, 2015) in R. To avoid convergence failure, the BOBYQA algorithm was used as the optimizer, as suggested by Bolker (2014).

The unconditional model was called Model 1. In Model 2, morpheme was added to the fixed-effects part because it was the predictor that most dramatically decreased the AIC. A comparison of Model 1 with Model 2 tested whether different morphemes were at different accuracy levels. In Model 3, I added by-morpheme random contrasts to Model 2. A comparison of Model 2 with Model 3 tested whether it was worth allowing the accuracy difference between morphemes to vary across learners. Likewise, in Model 4 and Model 5, I tested the effects of L1type and writingnum, respectively. With Model 6, I further added the by-writingnum random slope. In Models 7 through 9, I examined the effects of proficiency, morpheme × proficiency interaction, and L1type × writingnum interaction, respectively.

Table 2 shows the summary of model comparisons. The first three columns give the model number and the variables included in the fixed and random effects of the model. The fourth column lists the AICs of the model, and the fifth column shows the difference of AIC in comparison to the previous model. A negative value means that this model has better predictive accuracy than the model that precedes it. The last two columns show the results of likelihood ratio tests comparing the model with the previous model. The table indicates that AIC categorically decreased until Model 9, and likelihood ratio tests similarly suggested steady improvement until Model 9. No other term (e.g., morpheme × writingnum interaction) further decreased AIC.

TABLE 2

Although Model 9 may appear to be the most plausible model, Model 8 was selected as the final model because the decrease of AIC from Model 8 to Model 9 (–2.8) was fairly small and the *p* value of the added parameter (.029 for the L1type × writingnum interaction) was also not as low as the *p* values of other parameters. Given that both models were already highly

complex and that AIC tends to prefer more complex models in general (Held & Bové, 2014), I opted for the less complex model. As a reference, I also constructed a model that had the same structure as Model 8 but that did not include L1type, proficiency, or any interaction terms involving them. A comparison between Model 8 and this reference model indicated the extent to which L1 type and proficiency explain the variance.

The forward-selection procedure employed above may result in underspecified models because there can be a model that is better than the final model and includes a combination of parameters untested in the model selection procedure. To mitigate the potential effect of the procedure, a series of models were built in the following manner.[2] Instead of adding one variable at a time, I added two variables that sequentially decreased AIC the most. I then deleted one variable that resulted in the minimum increase in AIC. This procedure was repeated until no iteration decreased AIC further. This process partially avoided the potential underspecification issue because the procedure allowed for the exploration of part of the parameter combination space that was not tested in the pure forward-selection procedure. This 2-in-1-out procedure resulted in Model 9 as the final model, thereby partially confirming the robustness of my model. For the reason described above, however, I took Model 8 as the final model.

Interpretation of the Model

Interpreting Random Effects

Table 3 presents the random-effects components of the mixed-effects model. It also shows the random effects of the Reference Model, against which the effects of predictors in Model 8 were tested. In Table 3, the intercept rows represent the standard deviation of random intercepts for L1 and learner, and the other rows show the standard deviation of by-morpheme random contrasts and by-writingnum random slopes.

TABLE 3

The Reference Model indicates that the standard deviation of the by-L1 random intercept (Row 2 in the table) was .300, which indicates the dispersion of L1 groups in absolute accuracy in the logit scale. Similarly, the by-learner random intercept (Row 4) was .495, which is the magnitude of individual differences in article accuracy within each L1 group after progressing the mean number of writings (i.e., standardized writingnum = 0). The standard deviations of the by-morpheme random contrasts (Row 5-7) was .716 for past tense *–ed* and .582 for plural *–s*, and denote individual differences in the accuracy difference between articles and the morphemes. The standard deviation of the by-writingnum random slope (Row 8) was .192, which represents the magnitude of individual variation in the overall learning rate. When the values in Model 8 were examined, a fair amount of decrease in the by-learner random intercept (.495 → .412, or –20.0%) was seen. This shows the extent to which learners' overall proficiency and L1type explain individual variation in article accuracy. The by-morpheme random contrast similarly decreased in Model 8 (.716 → .613, or –16.8%, for past tense *–ed* and .582 → .481, or –21.1%, for plural *–s*). This represents the degree to which proficiency (but not L1type due to the absence of L1type-morpheme interaction in the fixed-effects structure) explains individual variation in between-morpheme accuracy difference.

Surprisingly, the by-writingnum random slope increased from the Reference Model to Model 8 (.192 → .197, or +2.3%). This is rooted in the fact that some of the within-learner variance could be reflected as between-learner variance in mixed-effects modeling (Hox, 2002; Snijders & Bosker, 1994). As a result, a within-learner predictor might explain both within- and between-learner variance. Maximum likelihood estimation (MLE) generally corrects this. However, when a predictor is centered or standardized within learners as in the present case, it

results in smaller between-learner variation in the average predictor value than that which is embedded in the correcting mechanism of MLE. This invites overcorrection by MLE, and random effects may increase as a result. Unlike linear mixed-effects models where it is possible to analytically compute the parameter values that maximize likelihood functions, in GLMMs we can only numerically approximate them (Bolker et al., 2009). The idea is the same, however. This, therefore, does not mean model misspecification.

Because the value is larger in past tense *–ed* random contrast than in plural *–s* random contrast (see Table 3), a larger individual variation remained in the accuracy difference between articles and past tense *–ed* than in the accuracy difference between articles and plural *–s*. However. because between- and within-learner variability is not completely independently quantified even in mixed effects models, random-effects components of different models are not strictly comparable. Comparison, however, is a common practice (e.g., Hox, 2002) and is still a useful strategy by which to examine the effect of predictors on random-effects components.

Interpreting Fixed Effects

I now turn to fixed-effects (Table 4); *p* values indicated by asterisks are only approximate, but a parametric bootstrap—a resampling technique that compares the target model with the reduced model that does not include the interested parameters (Pinheiro & Bates, 2000)—agreed with the significance of parameters in the table with the significance level of $p < .05$ based on 1,000 samples.

TABLE 4

Thus, the following observations can be made about the results summarized in Table 4.

- The main effect of morpheme (Rows 2 to 4 in the table) is significant. At the mean proficiency level, the accuracy of plural *–s* is generally higher than that of articles.

- The main effect of L1type (Rows 5 to 6) is also significant. The PRESENT group overall outperformed the ABSENT group at standardized writingnum = 0.
- The main effect of writing number (Row 7) is significant. As learners wrote, morpheme accuracy increased.
- The main effect of proficiency (Row 8) is significant and positive. Article accuracy tended to be higher in learners of higher proficiency.
- The morpheme × proficiency interaction (Rows 9 to 11) shows that the accuracy increase over proficiency was smaller in plural *–s* than in articles. In plural *–s*, accuracy increase per standard deviation of proficiency was nearly negligible (.238 –.224 = .015).[3]

Some of the terms that were not significant included the following:

- The L1type × proficiency and L1type × writingnum interactions are not present in the final model. This means that there is no evidence showing different cross-sectional or longitudinal developmental patterns between the PRESENT and ABSENT groups. This is interesting because the PRESENT group generally outperformed the ABSENT group, and the members of this group could have been more likely to approach the ceiling level of performance.
- The morpheme × writingnum interaction was not retained in the final model. This outcome shows that the rate of longitudinal development is similar across morphemes.

To look into the magnitude of individual variation, it is interesting to compare random effects in Table 3 with the corresponding fixed effects in Table 4. The fact that the random contrast for past tense *–ed* is .613 and its estimate in the fixed-effects structure is .141 means that at the mean proficiency level, the standard deviation of individual variation in the accuracy difference between articles and past tense *–ed* is much larger than the mean accuracy difference

between the two morphemes, which in turn indicates that although past tense *–ed* is more accurate than articles on average in this sample, the accuracy order between the two morphemes depends heavily on learners. The case of the by-writingnum random slope is similar. The standard deviation in Table 3 is .197, while the coefficient in the fixed effects is .082. This indicates that while on average learners' longitudinal development is characterized by increased accuracy, for a great proportion of learners, accuracy decreased overall. This is not the case for the difference between articles and plural *–s*, however. Because its random slope (.481) is smaller than the fixed-effects coefficient (.787), plural *–s* was usually (though not necessarily always) more accurate than articles in individual learners. At higher proficiency levels, however, the mean difference between the two morphemes decreased, as reflected in the negative coefficient of the interaction between proficiency and plural *–s*. The proportion of the learners whose accuracy was higher in articles than in plural *–s* was expected to increase. The discussion here illustrates that it is possible to quantify individual variation through GLMMs.

Summary of the GLMM Approach

In this section, I demonstrated systematicity (e.g., plural *–s* is on average more accurate than articles) and individual variation in the L2 accuracy of grammatical morphemes. In addition to its ability to model systematicity and individuality simultaneously, a particular strength of the GLMM is its feature of quantifying individual variation through random effects. Variance in random effects is informative as to (a) the extent to which individual variation is present in a certain effect (e.g., the standard deviation of the individual variation in article accuracy is .412 in logit scale), (b) whether it is larger or smaller compared to individual variation in another effect (e.g., individual variation in accuracy difference between articles and past tense *–ed* is larger than the variation in the difference between articles and plural *–s*), and (c) the degree to which

predictors explain variation (e.g., proficiency decreases the accuracy difference between articles and plural *–s* by 21.1%).[4]

Accounting for Nonlinearity and Individuality: Generalized Additive Mixed Models

In the previous section, the analysis assumed a linear change of accuracy in both cross-sectional and longitudinal development. The assumption, however, is unwarranted, particularly in light of prior SLA research demonstrating nonlinear learning curves (DeKeyser, 1997; Lightbown, 1983). This section examines whether the developmental path varies depending on learners' L1 types and morphemes when nonlinear development is assumed.

Brief Overview of Generalized Additive Models

Generalized additive models (GAMs) extend generalized linear models (GLMs) by modeling nonlinear relationships between independent and dependent variables. They achieve nonlinearity through the use of *splines*. The following explanation of splines is largely based on James et al. (2013) and Hastie, Tibshirani, and Friedman (2009).

A traditional way of modeling nonlinearity is by using polynomial functions. However, they cannot model flexible shapes without spending a large number of degrees of freedom, and doing so renders the resulting model unstable. In regression splines, one polynomial function models only part of the data, and multiple functions are used to cover the entire data. These functions are smoothly connected so that there is no wide jump in the predicted value. This point is illustrated in the upper two panels in Figure 3. Figure 3A demonstrates morpheme development in hypothetical learners. The dashed line represents the predicted values of accuracy based on a cubic function of proficiency (i.e.,$TLU = \beta_0 + \beta_1 \times$ proficiency $+ \beta_2 \times$ proficiency$^2 + \beta_3 \times$ proficiency^3, where βs are estimated from the data).[5] Here, we observe relatively large differences between observed (i.e., small circles) and fitted (i.e., dashed line)

values. A cubic function is thus inadequate for modeling accuracy development in this dataset. The solid line is a piecewise cubic function. Datapoints were horizontally divided into five equally spaced regions, and a cubic function was fitted to each region. We see that the predicted function is absurd as a whole: The lines are not connected and there are jumps in the fitted value as a result. Thus, simply employing multiple piecewise polynomial functions is insufficient for modeling nonlinearity.

FIGURE 3

To achieve more natural modeling of nonlinearity, certain constraints can be imposed on the piecewise polynomial functions. Specifically, it is common to constrain piecewise cubic functions so that the values of the function and its first and second derivatives are continuous at *knots*, the points at which cubic pieces connect. This way, the function is not only continuous throughout but also smooth at the knots (Zuur, Ieno, Walker, Saveliev, & Smith, 2009). In Figure 3B, the same data points are modeled by a *smoothing spline*. Though conceptually somewhat different, it is mathematically a variant of the cubic splines discussed above. Based largely on cubic functions, the smoothing spline models the data well in the present case.

The spline balances difference between fitted and observed values and the roughness or wiggliness of the curve. If it is allowed to be infinitely wiggly, it goes through all of the observed data points and would clearly overfit the data by modeling noise in addition to the underlying shape, thereby making it difficult to generalize to new datasets. If, on the other hand, the spline is not allowed to be wiggly at all, it would end up being a straight line that models nonlinearity poorly. The smoothing spline achieves this bias-variance trade-off through a procedure called generalized cross validation, which is an approximation of leave-one-out cross validation commonly employed to evaluate statistical models. Conceptually, a spline function with a certain

degree of smoothness is fitted to all but one data point, and then the difference between the observed value of the omitted data point and its predicted value based on the spline (i.e., error) is calculated. This process is repeated as many times as there are data points. The average error in this procedure indexes the goodness of the degree of smoothness. This whole process is then repeated for a wide range of smoothness values, and the optimal wiggliness is found in which the average error is minimized (Wood, 2009; Zuur et al., 2009).

GAMs are a semi-parametric technique that combines the smooths discussed above with parametric terms, thereby allowing a statistical test of the significance of some terms while controlling for the nonlinear effects of other terms. The lower two panels of Figure 3 illustrate the importance of accounting for nonlinearity through GAMs. These two figures show hypothetical accuracy development in two L1 groups: L1 Japanese and L1 Spanish. Both groups show clear U-shaped developmental patterns, but the data for L1 Spanish learners were generated to mark higher accuracy overall than those for L1 Japanese learners throughout development. In Figure 3C, the pattern is modeled by a linear function. It forces linearity on the nonlinear shape, resulting in large differences between observed and predicted values. Because the model hardly explains variance and the residuals are large, the accuracy difference between the two L1 groups is nonsignificant, $t(197) = 1.643$, $p = .102$, when proficiency is (mis)controlled for, despite the consistently higher accuracy of the L1 Spanish learners. Figure 3D models the same data with a GAM based on a *thin plate regression spline* (Wood, 2003), an approximation to a thin plate spline (Wood, 2010), which is a generalized form of the cubic spline discussed earlier. The model was constructed with the mgcv package (version 1.8-6; Wood, 2006) in R. Here, without pre-specifying shape, the GAM accurately models the U-shape. This in turn results in much smaller residuals than in Figure 3C, and this time, the effect of L1 is

correctly identified, $t(191.7) = 7.858$, $p < .001$. Therefore, the GAM was able to model the usual parametric term (L1) and nonparametric smooth (nonlinear effect of proficiency) simultaneously.

An exciting recent development is the incorporation of random effects into GAMs, making the model capable of accounting for nonlinear patterns of individual learners. The model, referred to as a generalized additive mixed model (GAMM; Baayen, 2014a, Chapter 8, 2014b; Wood, 2004, 2006), can construct separate wiggly curves for each learner by *penalized factor smooths*, which achieve the interaction between smooths and factors with the same degree of smoothness across learners (Wood, 2014). GAMMs have been used in psycholinguistics (e.g., Balling & Baayen, 2012; Mulder, Dijkstra, Schreuder, & Baayen, 2014), sociolinguistics (Wieling, Montemagni, Nerbonne, & Baayen, 2014), and SLA (Ning, Shih, & Loucks, 2014).

Model Specification and Model Selection

Models assumed binomial error distribution and employed a logit link function. The dependent variable and the potential independent variables were the same as the GLMM's, except that nonlinear terms were also considered. The interaction between two nonlinear terms (i.e., proficiency × writingnum) was entered as a *tensor product smooth*. Tensor product smooths extend nonlinearity to more than one dimension and model wiggly surfaces between the variables of naturally different scales (Hastie et al., 2009; Wood, 2010). A separate smooth was constructed for each factor level when L1 type or morpheme interacted with proficiency and/or writingnum. For example, in the specification of L1type × proficiency interaction, separate proficiency curves were created for each L1 type. Thus, unlike interactions in typical regression models, factor-smooth interactions in GAMs also account for the main effects of the continuous variables included in the interaction. Due to a centering constraint, factors need to be specified in the model separately.

Due to the high computational cost of GAMMs, building a model takes a relatively long time, and it was impractical to run, in model selection, the forward selection process that requires building multiple models at each step. Instead, I started with a model that was conceptually equivalent to the final GLMM constructed earlier and tested whether all of the parameters included in the model were necessary and whether including additional terms improved the model. Model 1, thus, included the following terms:

- L1type and morpheme as fixed effects,
- by-L1 random intercepts, by-learner random intercepts, and by-morpheme random contrasts at the level of individual learners as random effects,
- (standardized) writingnum and by-morpheme (standardized) proficiency as smooth terms to capture their potentially nonlinear effects, and
- by-writingnum random wiggly curves at the learner level.

Smooth terms were specified with thin plate regression splines. Random wiggly curves are similar to random slopes but also allow nonlinearity in the longitudinal developmental patterns of individual learners. Maximum likelihood estimation was employed. This model was different from the final GLMM in that nonlinear effects were assumed in proficiency and writingnum, and random wiggly curves were assumed instead of random slopes for individual learners.

With this model as the starting point, I first tested whether any additional terms improved the model. For this purpose, five candidate models were built.

1. Model 1 + L1type × morpheme interaction in the fixed-effects structure (B = –.004, p = .985 for PRESENT – past tense *–ed*; B = –.125, p = .355 for PRESENT – plural *–s*).
2. Model 1 – writingnum smooth + writingnum smooth for each morpheme (i.e., writingnum × morpheme interaction (χ^2 = .115, p = .735 for the writingnum curve for

articles; $\chi^2 < .001$, $p = .999$ for the writingnum curve for past tense *–ed*; $\chi^2 = .011$, $p = .915$ for the writingnum curve for plural *–s*).

3. Model 1 + proficiency smooth for each L1 type ($\chi^2 < .001$, $p = .999$ for the proficiency curve for the ABSENT group; $\chi^2 = .029$, $p = .864$ for the proficiency curve for the PRESENT group)
4. Model 1 + proficiency × writingnum interaction realized as a tensor-product interaction ($\chi^2 = 5.327$, $p = .419$).
5. Model 1 - writingnum smooth + writingnum smooth for each L1type (i.e., writingnum × L1type interaction; $\chi^2 = .027$, $p = .871$ for the writingnum curve for the ABSENT group; $\chi^2 = 19.416$, $p < .001$ for the writingnum curve for the PRESENT group).

The model selection procedure, based on *p* values in the parentheses (Wood, 2013a, b), suggests that Candidate Model 5 is better than Model 1, and AIC-based model comparison supports the decision as well ($\Delta AIC = -8.2$). This model is referred to as Model 2.

The next step was to test whether it was worth adding further terms. The same procedure was again followed, except that candidate terms were added to Model 2 this time. The added terms were the same as Candidate Models 1 through 4 above. The process indicated that none of the terms improved the model ($p > .105$ for all of the terms). I then examined whether all of the terms in Model 2 were needed. The *p* values of Model 2 parameters indicate that while some parameters are nonsignificant (e.g., $\chi^2 = 2.668$, $p = .102$ for the proficiency curve for past tense *–ed*), they are restricted to the levels of the factors or the levels of the interaction terms involving the factors whose other levels are significant (e.g., $\chi^2 = 21.869$, $p < .001$ for the proficiency curve for articles). This indicates that all of the terms should be kept in the model.

Model 2, however, suggests that the effect of proficiency is linear (EDF = 1.000 for all of

the morphemes). The proficiency term, therefore, was moved to the parametric part: Model 3 included L1type, proficiency, morpheme, and the proficiency × morpheme interaction as fixed-effects parametric terms. This did not affect AIC (ΔAIC = –.002). To further test whether random wiggly curves were necessary, another model was constructed in which random wiggly curves in Model 3 were replaced with by-writingnum random slopes. In other words, the model assumed linear effects of writingnum at the level of individual learners. Model comparison indicated that random wiggly curves needed to be included (ΔAIC = 112.5), suggesting that the learning curve was nonlinear at the level of individual learners.

The above did not directly indicate whether separate writingnum curves were needed for the two L1 types. To analyze this, a separate curve was estimated on top of the curve for the reference level. In other words, to examine whether L1 type affects the longitudinal developmental pattern, two separate curves were constructed: one for the ABSENT learners and the other for the PRESENT learners on top of the curve for the ABSENT group (Baayen, 2014a, Chapter 8; Wieling, 2015; Wood, 2014). If the latter proved to be significant, it would suggest that it is worth having an additional curve for the PRESENT group on top of the ABSENT group curve, which in turn means that the longitudinal developmental pattern differs across L1 types. The results suggested that a separate writingnum curve was needed for the PRESENT group (χ^2 = 13.472, p = .012). Model 3 was thus the final model; it is explored below.

Interpretation of the Model

Tables 5 through 7 show the results of the final model. Parametric terms (Table 5) suggest that (a) PRESENT learners generally outperformed ABSENT learners (Row 3), (b) higher proficiency learners tended to be more accurate in using articles than lower proficiency learners (Row 4), (c) learners were more accurate in the use of plural *–s* than articles at the mean

proficiency level (Row 7), and (d) cross-sectional accuracy increase was smaller in plural *–s* than in articles (Row 10). The nonsignificance of L1type × proficiency interaction indicates that the cross-sectional developmental pattern can be assumed to be similar across L1 types.

TABLE 5

Table 6 shows estimated degrees of freedom (EDF), reference degrees of freedom (Ref.df), χ^2, and *p* values for the splines. If the EDF is close to 1 as it was in the effect of writingnum in the ABSENT group, the relationship between independent and dependent variables is close to linear in logit scale (Baayen, 2010a), and the larger its value, the wigglier the curve is. The table shows linearity in the partial effect of writingnum for the ABSENT group (EDF = 1.001 in Row 2) but nonlinearity for the PRESENT group (EDF = 3.503 in Row 3). The table also indicates significant individual variation in longitudinal development (Row 4).

TABLE 6

Table 7 provides the standard deviation of random effects.[6] As in the GLMM, between-L1 variation in absolute accuracy and individual variation in the accuracy difference between articles and other morphemes can be observed. Drawing inferences from the above tables, however, is not necessarily straightforward: Smooth terms in Table 6 make interpretation especially difficult. I turn now to one strategy that can assist in drawing inferences from the results: visualizing the fitted values.

TABLE 7

Figure 4 shows the fitted nonlinear accuracy development in individual learners. The upper panel represents adjustments to logit TLU scores for individual learners across standardized writing numbers. If there were no individual variation within each L1 type, morpheme, and proficiency level, all of the lines should completely overlap. As we can see,

however, large individual variation is present both in terms of absolute accuracy and developmental shape. The figure demonstrates large individual variation well, but it does not show how learners develop in the scale of TLU scores in a particular morpheme. The bottom four panels in Figure 4, therefore, show the fitted values of article accuracy in individual learners divided into two proficiency groups (higher vs. lower) and two L1 types (ABSENT vs. PRESENT). The cut-off proficiency level for the two proficiency groups was learners' mean proficiency. The thick lines in each panel are locally weighted scatterplot smoothing lines (LOESS; Larson-Hall & Herrington, 2010; Singer & Willett, 2003) showing the overall trend. Although the parametric terms in Table 5 indicate that the PRESENT group outperformed the ABSENT group, this is hardly visible in Figure 4 due to individual variation within each L1 type. Furthermore, whereas, on average, higher proficiency learners used articles more accurately than lower proficiency learners, this was merely a tendency and only characterized the development of the hypothetical "average" learner. Individual variation definitely outweighs the typological difference in L1 and can also have a larger impact than general proficiency. Moreover, the developmental pattern slightly differs between the ABSENT and PRESENT groups, as Table 6 indicates. However, the figure also suggests that this difference is marginal compared to the scale of individual variation.

FIGURE 4

Summary of the GAMM Approach

The GAMM took into account individual variation and nonlinearity and modeled accuracy development as a function of proficiency, longitudinal development, and L1 type. The final model demonstrated (a) individual variation in absolute accuracy and in nonlinear development, (b) systematic L1 influence and proficiency effects on absolute accuracy, and (c) L1 influence on

longitudinal developmental patterns. The empirical and quantified demonstration of nonlinearity, individual variation, and systematicity was only achievable through GAMMs.

Contrasting GLMM/GAMM with GLM/GAM

Now that both types of models have been explored, they are compared against each other and against GLMs and GAMs, the models that do not account for individual variation. The only difference between the GLMM and the GAMM is that the GAMM includes the L1type × writingnum interaction while the GLMM does not. Recall that the term was at the borderline in the model selection process of GLMMs. Figure 4, based on the final GAMM, also shows that the difference in the developmental curve between the two L1 types is minute, especially in view of large individual variation. Thus, although it is worth including the interaction term in the model when nonlinearity is accounted for, I conclude that its effect is nearly negligible from a practical perspective.

It is also interesting to compare GLMM/GAMM with GLM/GAM because such a comparison highlights the importance of taking individual variation into account when modeling L2 development. As in the GLMM, GLMs and GAMs were constructed based on the forward selection approach. Both the GLMs and the GAMs used the logit link function and assumed binomial error distribution. The final GLM included morpheme, L1 type, proficiency, writingnum, morpheme × proficiency interaction, and proficiency × L1type interaction. The final GAM included morpheme and L1 type as parametric terms, and as smooth terms separate wiggly proficiency curves for each morpheme, separate writingnum curves across L1 types, and a proficiency × writingnum wiggly surface. The detailed model selection procedure is provided in Appendix S4 in the Supporting Information online.

The results showed a few conflicting findings between GLMM/GAMM and GLM/GAM.

More specifically, the GLM supported the L1type × proficiency interaction and the GAM included the proficiency × writingnum interaction, while GLMM and GAMM supported neither. In addition, the GAM suggested nonlinear cross-sectional development, while the GAMM demonstrated linear development. The findings of the GLMM/GAMM were more conservative than those of the GLM/GAM: The GLM/GAM either pointed toward more significant parameters than the GLMM/GAMM or suggested nonlinear effects when the GAMM indicated linear effects. These are all likely to be rooted in whether individual variation is taken into account (GLMM and GAMM) or not (GLM and GAM). Generally speaking, ignoring the nested structure of data results in unfairly small standard errors (Hox, 2002; Long, 2012), leading to narrower confidence intervals (cf. McKeown & Sneddon, 2014; Wieling, 2015). In the present context, because the GLM and the GAM ignore the dependency of data within individual learners, their standard errors turned out to be unfairly small, inviting spurious significant results.

The difference between the models is illustrated in Figure 5, which shows the predicted cross-sectional and longitudinal development of article accuracy in two learners, one L1 Russian and one L1 Brazilian, who contributed the largest number of error-tagged writings among the ABSENT and PRESENT learners, respectively. The point of the figure is the magnitude of uncertainty represented by the width of shaded 95% confidence intervals, which are clearly wider in the GLMM and the GAMM panels than in the GLM and the GAM panels. The wider confidence intervals of the GLMM and GAMM are brought about by their ability to account for individual variation. The GLM suggested that cross-sectional developmental patterns varied across L1 types because models were (erroneously) certain of the trajectory of each L1 type and the trajectories differed, while the GLMM and the GAMM were much less certain that the two trajectories were different. Similarly, the GAM judged cross-sectional developmental patterns to

be nonlinear because the narrow confidence intervals and a relatively fixed trajectory that resulted suggested this, while the wide confidence intervals of the GAMM and the resulting uncertainty in the trajectory did not support it. Thus, the GLMM/GAMM results are more trustworthy, and the illustration here demonstrates the significance of accounting for individual variation in modeling L2 development.

FIGURE 5

Discussion

GLMMs and GAMMs in this article demonstrate nonlinearity and individual variation in the L2 development of English grammatical morphemes. SLA researchers have shown interest in these phenomena but were previously unequipped with the analytical tools to investigate them. With sophisticated statistical models of the type employed in this paper, however, complex phenomena such as L2 development can be modeled, allowing much less information to be lost than when traditional statistical techniques are used.

More specifically, this article shows that (a) plural *–s* was more accurate than articles in general, (b) learners with an equivalent feature in their L1 outperformed those whose L1s lack the feature, (c) article accuracy increased as learners' proficiency rose, (d) cross-sectional developmental patterns varied across morphemes, and (e) large individual variation was present in absolute accuracy, the accuracy difference between morphemes, and longitudinal developmental patterns. There was no disagreement in the above findings between GLMM and GAMM. Thus, these conclusions can safely be accepted.

The cross-sectional developmental patterns varied between articles and past tense *–ed* on the one hand and plural *–s* on the other. Articles and past tense *–ed* underwent more rapid increase in accuracy than did plural *–s*, whose accuracy remained relatively unchanged

throughout development. This difference was likely due to the higher accuracy of plural *–s* and, as a result, to the ceiling effect. It is interesting, however, that no significant difference was observed between the developmental patterns of articles and past tense *–ed* despite the fact that the article is a nominal free morpheme that encodes a language-dependent concept (i.e., definiteness) and past tense *–ed* is a verbal bound morpheme that encodes tense, a fairly language-independent concept. This finding shows that the classic distinctions between morphemes may not strongly influence the developmental trajectory of morpheme accuracy. Because this article targeted only three morphemes, this observation is merely suggestive rather than conclusive.

The current analyses also demonstrate systematicity, individuality, and nonlinearity in L2 development. L1 type consistently exerted influence on accuracy, demonstrating that accuracy is not determined randomly. However, as has been repeatedly emphasized throughout the paper, large individual variation was present as well, both in the absolute accuracy and in developmental patterns of morphemes. Together with the complex nonlinear patterns discussed earlier, I echo Baayen (2014b, p. 361):

> The results obtained with GAMs can be embarrassingly rich, in the sense that the results are far more complex than expected given current models. GAMs will often challenge the state of the art of current theories, and the author's intuition is that they may force the field to move more into the direction of dynamic systems approaches to language.

Although the claim was made in the context of GAMs and GAMMs, it fully applies also to GLMMs.

I now briefly summarize features of the models discussed in this paper and note their potential weaknesses. The defining property of GLMMs is that they incorporate both fixed- and

random-effects variables. This allows researchers to model systematicity and individuality simultaneously. GLMMs, however, are not very flexible in modeling nonlinearity. GAMMs can model nonlinearity and individual variation simultaneously. They cannot, however, currently handle a correlation parameter in random effects (Wieling, 2015). A further potential drawback is that they are less interpretable than other simpler models like GLMs/GLMMs (cf. James et al., 2013). It is worth noting, however, that interpretability of simpler models may come at the cost of less precision.

This article is not without its limitations. Accuracy was calculated by aggregating all error types. However, different mechanisms may operate between omission, misformation, and overgeneralization errors or between definite and indefinite article uses. Ideally, error type should be incorporated into the model. Additionally, models only included developmental measures (i.e., proficiency and writing number) and L1-related variables (i.e., L1 and L1 type) as predictors of accuracy. Many more variables are certain to affect accuracy, such as tasks and linguistic contexts. Further investigation into the sources of variability should shed light on why cross-sectional and longitudinal developmental patterns take the form they do. The dataset itself is also a source of limitations. For example, it is worth looking into the potential effects of other variables, including tasks, teaching materials in Englishtown, and varying progress rates across learners (Alexopoulou et al., 2015).

Conclusion

In this present article, I introduced statistical models that capture systematicity, individuality, and nonlinearity and illustrated their potential in SLA research with the L2 accuracy development of English grammatical morphemes as an example. In light of the nonlinear and variable nature of L2 development, these techniques help researchers to better model L2 development and provide

insights into the complex, dynamic, and nonlinear process of development.

Final revised version accepted 4 September 2015

Notes

1 An alternative approach is to use national language. See Alexopoulou, Geertzen, Korhonen, and Meurers (2015) for details.

2 I thank an anonymous reviewer for suggesting this procedure.

3 The calculation may not look correct due to rounding, but the value is accurate.

4 There are two further features of GLMMs that merit discussion but the space does not allow to elaborate: the correlation structure of random effects and shrinkage. They are demonstrated in Appendix S3 in the Supporting Information online.

5 Although accuracy is proportional, logistic regression was not employed in order to avoid confusion between linearity in probability scale and linearity in logit scale. The same follows for the remaining panels.

6 The values were calculated with the getSD.gam function in the paper package of Wieling et al. (2014), available at http://openscience.uni-leipzig.de/index.php/mr2/article/view/41.

References

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, *1*, 96–129. doi:10.1075/ijlcr.1.1.04ale

Atkinson, D. (2011). A sociocognitive approach to second language acquisition: How mind, body, and world work together in learning additional languages. In D. Atkinson (Ed.),

Alternative approaches to second language acquisition (pp. 143–166). Abingdon, UK: Routledge.

Baayen, R. H. (2010a). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*, 436–461. doi:10.1075/ml.5.3.10baa

Baayen, R. H. (2010b). A real experiment is a factorial experiment? *The Mental Lexicon*, *5*, 149–157. doi:10.1075/ml.5.1.06baaBaayen, R. H. (2014a). *Analyzing linguistic data: A practical introduction to statistics using R* (2nd edition). Manuscript in preparation.

Baayen, R. H. (2014b). Multivariate statistics. In R. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 337–372). Cambridge, UK: Cambridge University Press.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on Naive Discriminative Learning. *Psychological Review*, *118*, 438–481. doi:10.1037/a0023851

Baayen, R. H., Vasishth, S., Bates, D., & Kliegl, R. (2015). *Out of the cage of shadows*. Manuscript submitted for publication. Retrieved from http://arxiv.org/abs/1511.03120v1

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*, 80–106. doi:10.1016/j.cognition.2012.06.003

Barkaoui, K. (2014). Quantitative approaches for analyzing longitudinal data in second language research. *Annual Review of Applied Linguistics*, *34*, 65–101. doi:10.1017/S0267190514000105

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D. M. (2009). *[R-sig-ME] lmer: ML and REML estimation*. Retrieved from https://stat.ethz.ch/pipermail/r-sig-mixed-models/2009q1/002096.html

Bates, D. M. (2010). *Lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*. doi:10.18637/jss.v067.i01

Blevins, J. P. (2013). Word-based morphology from Aristotle to modern WP (Word and Paradigm models). In K. Allan (Ed.), *The oxford handbook of the history of linguistics* (pp. 375–395). Oxford, UK: Oxford University Press.

Bolker, B. M. (2014, January 26). *Convergence error for development version of lme4* [A response at a Q&A website]. Retrieved from http://stackoverflow.com/a/21370041/3237850

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*, 127–135. doi:10.1016/j.tree.2008.10.008

Brown, R. (1973). *A first language: The early stages*. London, UK: George Allen & Unwin.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J. (2010*). Language, usage and cognition*. Cambridge, UK: Cambridge University Press.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382. doi:10.1177/0267658312443651

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language

morphosyntax. *Studies in Second Language Acquisition*, *19*, 195–221. doi:10.1017/S0272263197002040

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, *82*, 39–54. doi:10.1111/1365-2656.12013

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford, UK: Oxford University Press.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning*, *23*, 245–258.

Ellis, N. C., & Schmidt, R. (1998). Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning or morphosyntax. *Language and Cognitive Processes*, *13*, 307–336. doi:10.1080/016909698386546

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 database: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Millar, K. I. Martin, C. M. Eddington, A. Henery, N. M. Miguel, A. Tseng, … D. Walter (Eds.), *Selected proceedings of the 2012 Second Language Research Forum. Building bridges between disciplines* (pp. 240–254). Cascadilla Proceedings Project.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Goldschneider, J., & DeKeyser, D. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50. doi:10.1111/1467-9922.00147

Gries, S. Th. (2013). *Statistics for linguistics with R: A practical introduction* (2nd edition).

Berlin, Germany: De Gruyter Mouton.

Gries, S. Th. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, *10*, 95–125. doi:10.3366/cor.2015.0068

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (second edition)*. New York, NY: Springer.

Hastie, T., & Tibshirani, R. J. (1990). *Generalized additive models*. London, UK: Chapman & Hall.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207. doi:10.3758/BF03212979

Held, L., & Bové, D. S. (2014*). Applied statistical inference: Likelihood and Bayes*. Heidelberg, Germany: Springer.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. New York, NY: Lawrence Erlbaum.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning*, *50*, 245–309. doi:10.1111/0023-8333.00118

Jarvis, S., & Pavlenko, A. (2007). *Crosslinguistic influence in language and cognition*. New York: Routledge.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*,

Article 238. doi:10.3389/fpsyg.2010.00238

Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning*, *61*, 1328–1354. doi:10.1111/j.1467-9922.2011.00638.x

Larsen-Freeman, D. E. (2011). A complexity theory approach to second language development/acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 48–72). Abingdon, UK: Routledge.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*, 368–390. doi:10.1093/applin/amp038

Lightbown, P. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. Seliger & M. H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 217–243). Rowley, MA: Newbury House.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207. doi:10.1111/lang.12117

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Thousand Oaks, CA: Sage Publications.

Luk, Z. P., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural *–s*, articles, and possessive *'s*. *Language Learning*, *59*, 721–754. doi:10.1111/j.1467-9922.2009.00524.x

Matuschek, H., Kliegl, R., Vasishth, S., Baayen., R. H., & Bates, D. (2015). *Balancing Type I error and power in linear mixed models*. Manuscript submitted for publication. Retrieved from http://arxiv.org/abs/1511.01864v1

McKeown, G. J., & Sneddon, I. (2014). Modeling continuous self-report measures of perceived

emotion using generalized additive mixed models. *Psychological Methods*, *19*, 155–174. doi:10.1037/a0034282

Mizumoto, A., & Plonsky, L. (2015). R as a lingua franca : Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*. Advance online publication. doi:10.1093/applin/amv025

Mulder, K., Dijkstra, T., Schreuder, R., & Baayen, H. R. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, *72*, 59–84. doi:10.1016/j.jml.2013.12.004

Murakami, A., & Alexopoulou, T. (2015). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*. Advance online publication. doi: 10.1017/S0272263115000352

Ning, L.-H., Shih, C., & Loucks, T. M. (2014). Mandarin tone learning in L2 adults: A test of perceptual and sensorimotor contributions. *Speech Communication*, *63-64*, 55–69. doi:10.1016/j.specom.2014.05.001

Odlin, T. (1989). *Language transfer*. Cambridge, UK: Cambridge University Press.

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition*, *6*, 69–78. doi:10.1017/S0272263100000309

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*, 445–485. doi:10.1080/01690960050119661

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.r-project.org

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Slobin, D. I. (1996). From "thought to language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge, UK: Cambridge University Press.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363. doi:10.1177/0049124194022003004

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*, 569–613. doi:10.1111/j.1467-9922.2010.00622.x

van Geert, P. (2008). The dynamic systems approach in the study of L1 and L2 acquisition: An introduction. *The Modern Language Journal*, *92*, 179–199. doi:10.1111/j.1540-4781.2008.00713.x

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior & Development*, *25*, 340–374. doi:10.1016/S0163-6383(02)00140-6

Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, *92*, 214–231. doi:10.1111/j.1540-4781.2008.00715.x

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*,

1182–9. doi:10.1111/j.1365-2656.2006.01141.x

Wieling, M. (2015). *Analyzing EEG data using GAMs: Lecture 4 of advanced regression for linguists*. Retrieved 16 January, 2015, from http://martijnwieling.nl/statscourse/lecture4/presentation.pdf

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen., R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*, *90*, 669–692.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*, e23613. doi:10.1371/journal.pone.0023613

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*, 95–114. doi:10.1111/1467-9868.00374

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686. doi:10.1198/016214504000000980

Wood, S. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Wood, S. (2009). GAMs: Semi-parametric GLMs. Retrieved from http://people.bath.ac.uk/sw283/mgcv/gam-theory.pdf

Wood, S. (2010). *A toolbox of smooths*. Retrieved from http://people.bath.ac.uk/sw283/mgcv/tampere/smooth-toolbox.pdf

Wood, S. (2013a). On *p*-values for smooth components of an extended generalized additive

model. *Biometrika*, *100*, 221–228. doi:10.1093/biomet/ass048

Wood, S. (2013b). A simple test for random effects in regression models. *Biometrika*, *100*, 1005–1010. doi:10.1093/biomet/ast038

Wood, S. (2014). *Package 'mgcv'*. Retrieved from http://cran.r-project.org/web/packages/mgcv/mgcv.pdf

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publishers's website:

Appendix S1. General Issues in Regression Modeling.

Appendix S2. Accuracy of the R Scripts Used to Identify Errors.

Appendix S3. Correlation Parameter and Shrinkage in Mixed-Effects Models.

Appendix S4. Generalized Linear Models and Generalized Additive Models.

Table 1 Descriptive statistics for target grammatical morphemes

	Nonzero observations		Obligatory contexts		Omissions errors		Overgeneralization errors	
Morpheme	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Articles	20.08	6.01	140.92	71.20	12.23	8.47	5.53	4.17
Past tense *–ed*	6.89	3.07	16.09	7.64	0.66	0.96	1.04	1.27
Plural *–s*	18.91	5.92	90.53	47.53	3.44	3.21	1.94	2.00

Table 2 Summary of GLMM model comparisons

	Model description				Test against prior model	
Model	Fixed-effects	Random-effects	AIC	ΔAIC	Statistic	*p*
Model 1	None	By-L1 + by-learner random-intercepts	13799.8			
Model 2	Model 1 + morpheme	Same as Model 1	13356.4	–443.4	$\chi^2(2) = 447.39$	< .001
Model 3	Same as Model 2	Model 1 + by-morpheme random-contrasts at learner level	13276.5	–80.0	$\chi^2(5) = 89.97$	< .001
Model 4	Model 2 + L1type	Same as Model 3	13253.7	–22.8	$\chi^2(1) = 24.80$	< .001
Model 5	Model 4 + writingnum (standardized)	Same as Model 3	13240.3	–13.4	$\chi^2(1) = 15.41$	< .001
Model 6	Same as Model 5	Model 3 + by-writingnum random-slope at learner level	13212.7	–27.5	$\chi^2(4) = 35.51$	< .001
Model 7	Model 5 + proficiency (standardized)	Same as Model 6	13197.9	–14.8	$\chi^2(1) = 16.81$	< .001
Model 8	Model 7 + morpheme × proficiency interaction	Same as Model 6	13188.3	–9.7	$\chi^2(2) = 13.67$	.001

Model 9	Model 8 + L1type × writingnum interaction	Same as Model 6	13185.5	–2.8	$\chi^2(1) = 4.75$	.029
Reference Model	Morpheme + writingnum (standardized)	Same as Model 6	13229.5			

Table 3 Random effects structure of GLMM Model 8 and reference model

	Factor	Random effects	*SD* in Model 8	*SD* in reference model
1	L1			
2		Intercept	0.295	0.300
3	Learner			
4		Intercept	0.412	0.495
5		Morpheme		
6		Past tense *–ed*	0.613	0.716
7		Plural *–s*	0.481	0.582
8		Writingnum (standardized)	0.197	0.192

Table 4 Fixed effects structure of GLMM Model 8

	Parameter	*B*	*SE*
1	Intercept	1.561**	0.123
2	Morpheme		
3	Past tense *–ed*	0.141	0.098
4	Plural *–s*	0.787**	0.063
5	L1type		
6	PRESENT	0.679**	0.123
7	Writingnum (standardized)	0.082*	0.027
8	Proficiency (standardized)	0.238**	0.043
9	Proficiency (standardized): Morpheme		
10	Proficiency (standardized): Past tense *–ed*	–0.115	0.089
11	Proficiency (standardized): Plural *–s*	–0.224**	0.059

Note. *$p < .01$, **$p < .001$.

Table 5 Parametric terms of GAMM Model 3

	Parameter	*B*	*SE*
1	Intercept	1.532***	0.127
2	L1type		
3	PRESENT	0.685***	0.118
4	Proficiency (standardized)	0.236***	0.050
5	Morpheme		
6	Past tense *–ed*	0.049	0.086
7	Plural *–s*	0.741***	0.062
8	Proficiency (standardized): Morpheme		
9	Proficiency (standardized): Past tense *–ed*	–0.100	0.084
10	Proficiency (standardized): Plural *–s*	–0.220***	0.059

Note. *$p < .01$, **$p < .001$.

Table 6 Smooth terms of GAMM Model 3

Term		EDF	Ref.df	χ^2	*p*
1	Writingnum (standardized): L1type				
2	Writingnum (standardized): ABSENT	1.001	1.002	0.026	.872
3	Writingnum (standardized): PRESENT	3.503	4.300	19.830	.001
4	By-writingnum random wiggly curve for individual learners	233.053	1415.000	867.258	.002

Table 7 Random effects of GAMM Model 3

Random effects	*SD*	*p*
By-L1 random intercepts	0.285	< 0.001
By-morpheme random slopes for individual learners	0.176	< 0.001

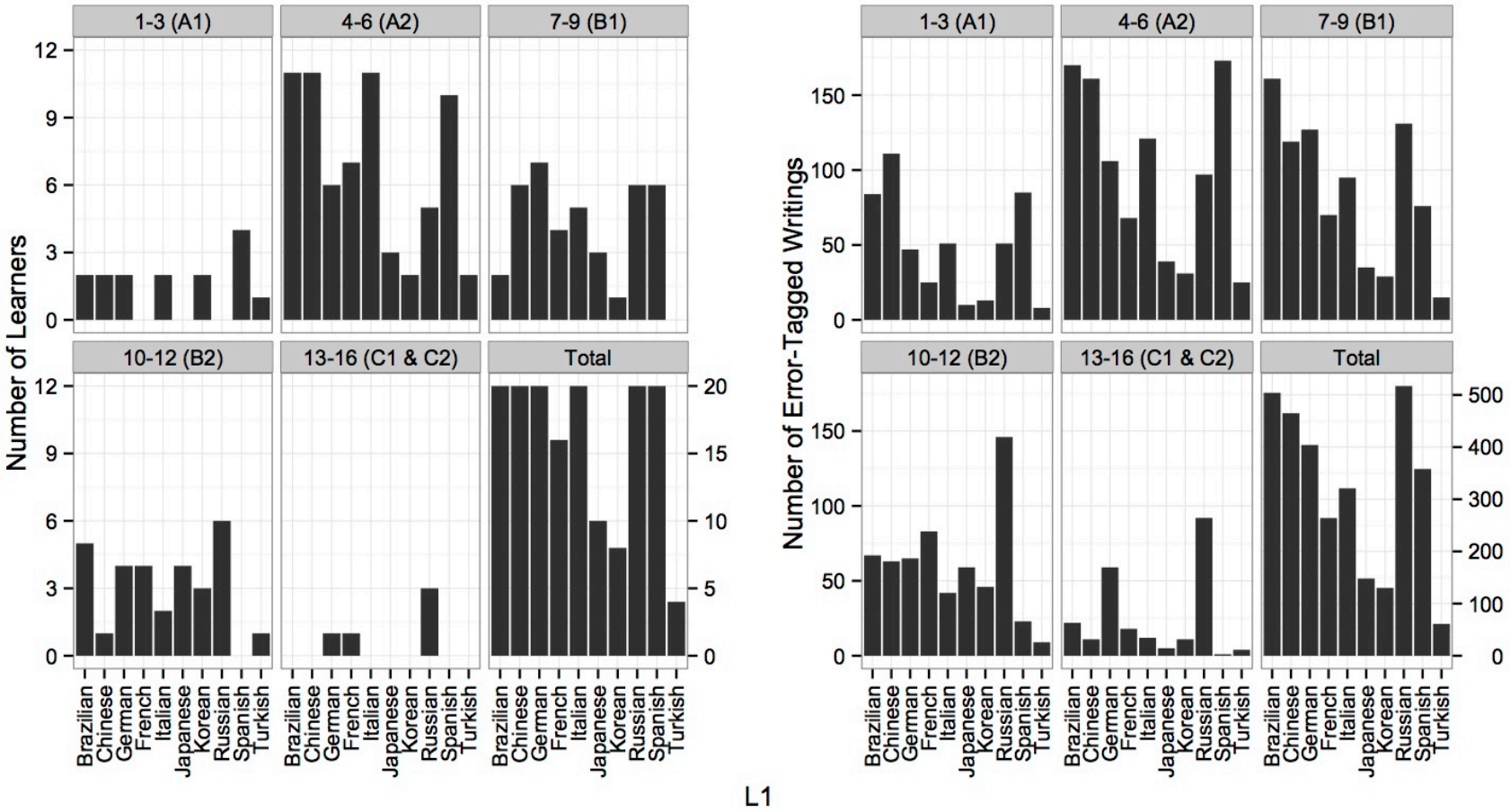

Figure 1 Number of learners and number of error-tagged writings in each L1 group at each CEFR level. The Total panel has a different y-axis scale from the other panels.

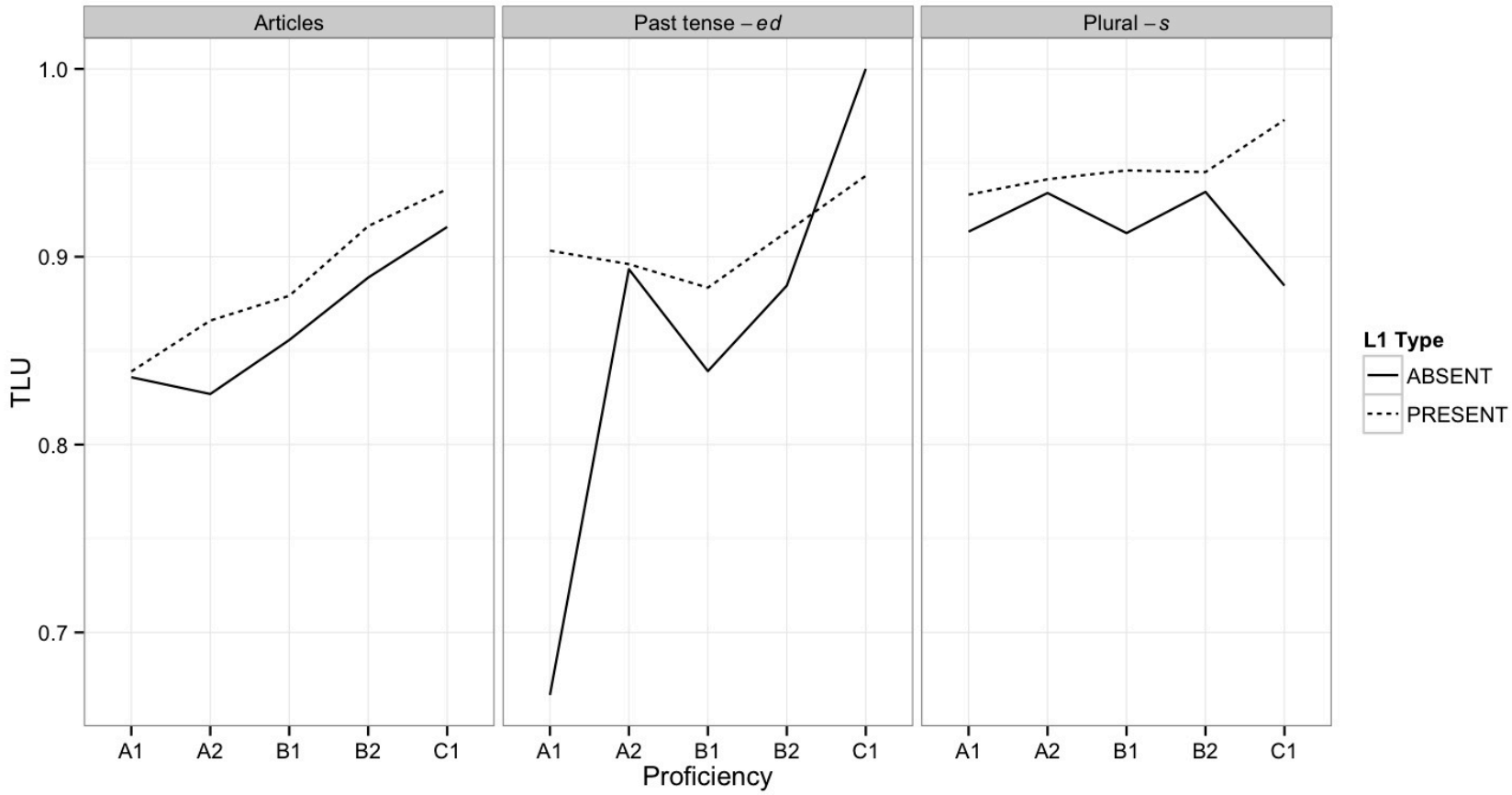

Figure 2 Cross-sectional development of morphemes across L1 types.

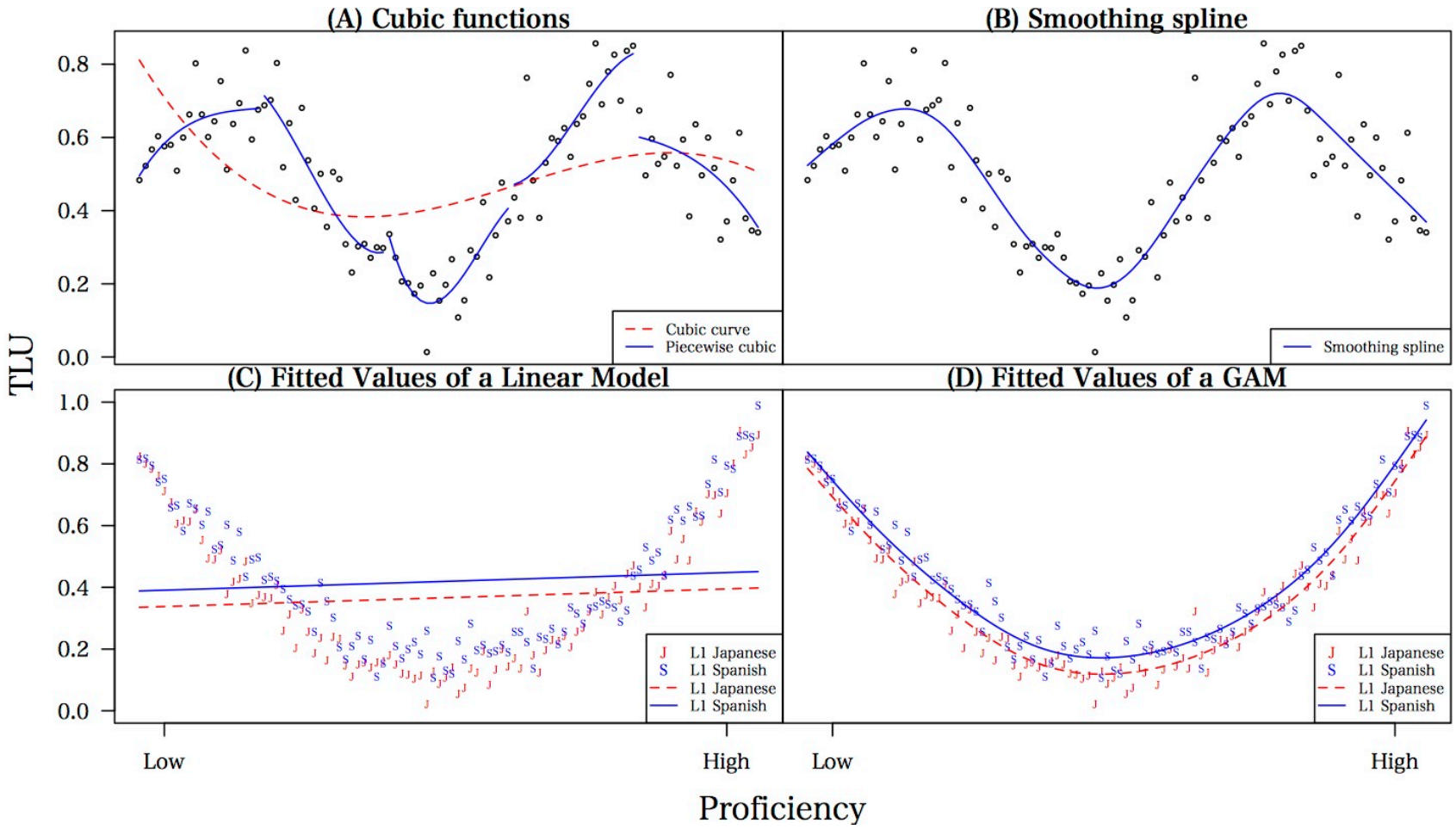

Figure 3 Illustration of splines and GAM.

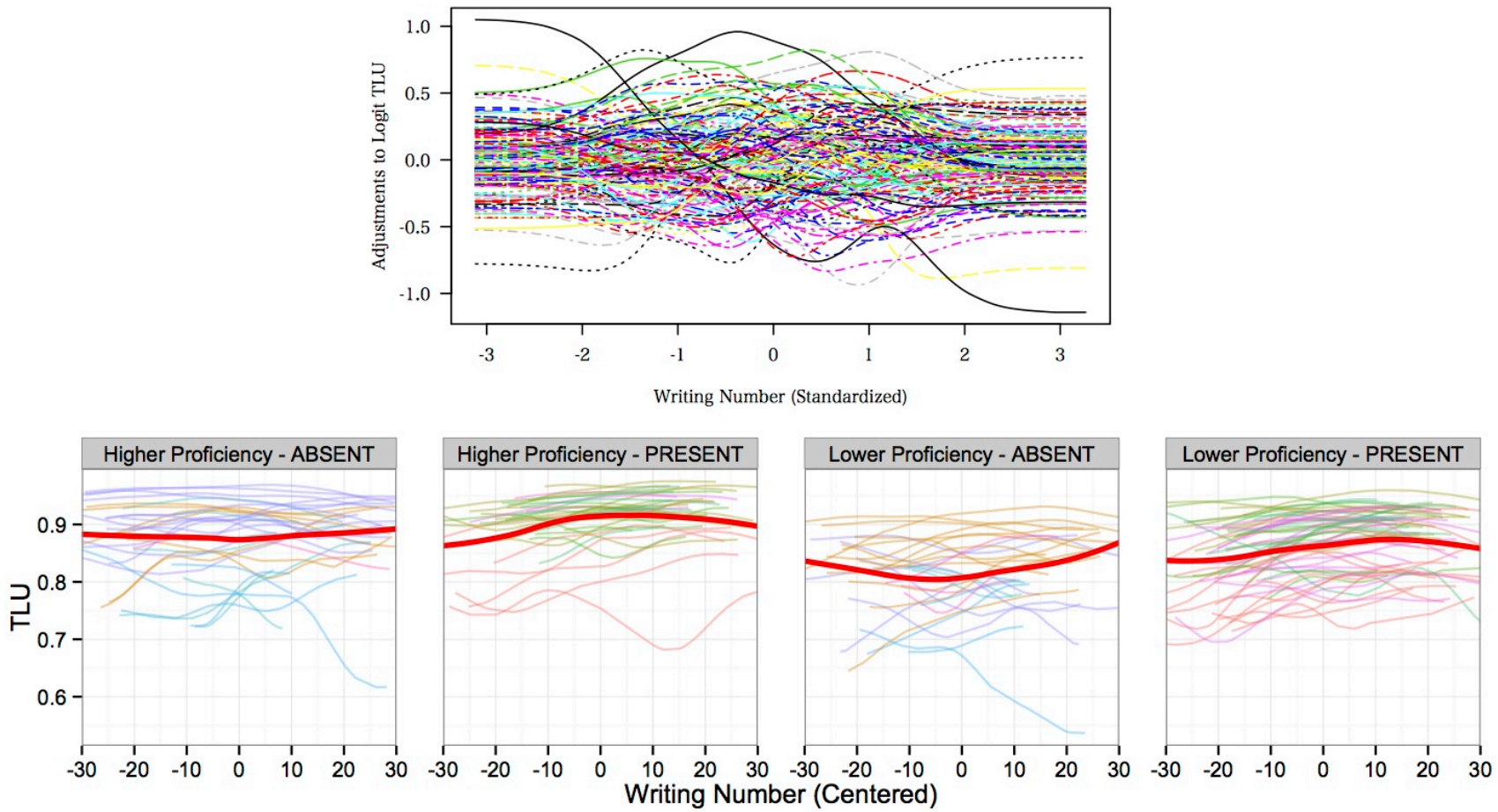

Figure 4 Individual variation in nonlinear longitudinal development.

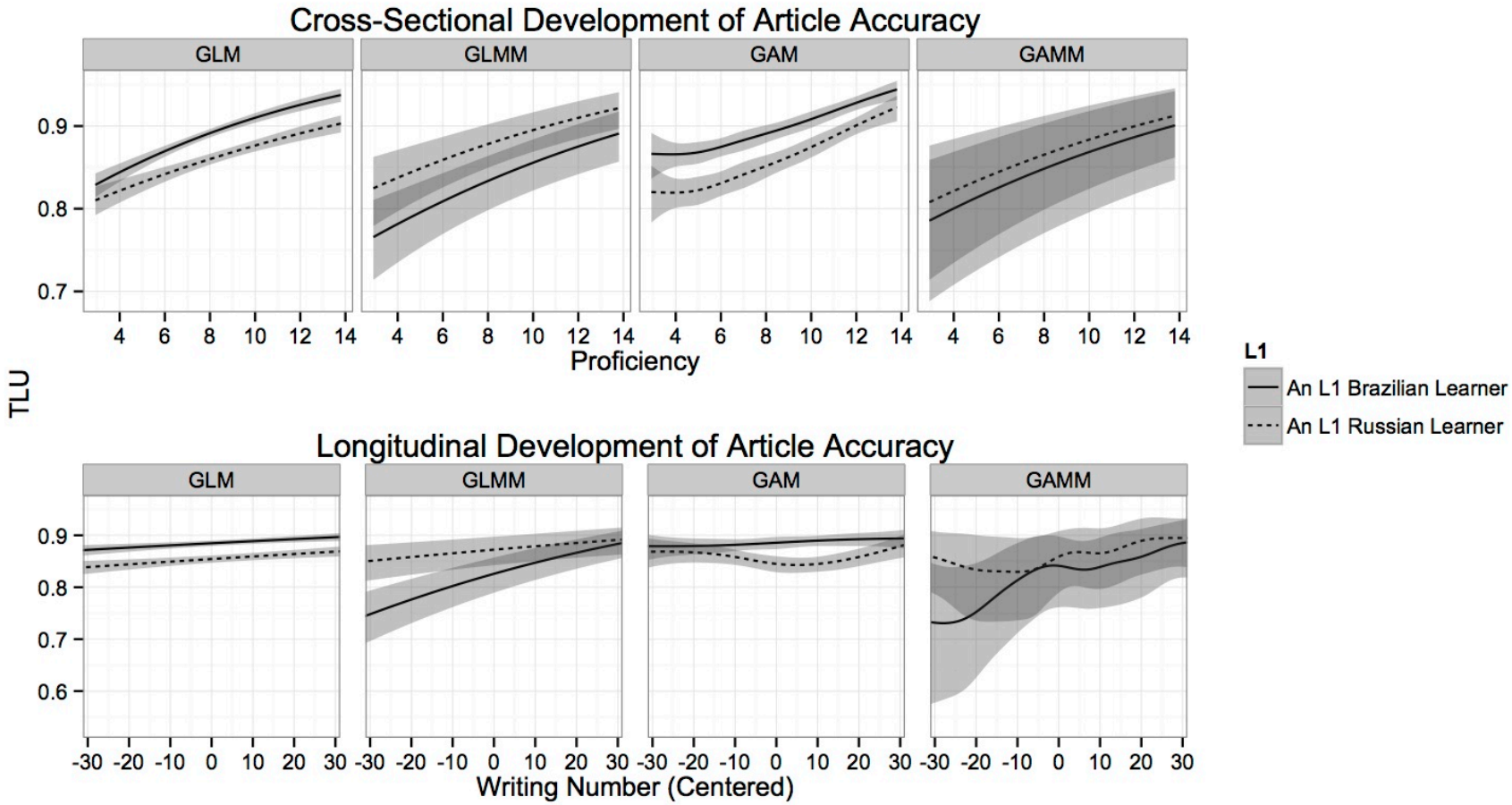

Figure 5 Cross-sectional and longitudinal development of article accuracy across different types of models for different L1 types.